Social Network Analysis and Mining (2019) 9:46
https://doi.org/10.1007/s13278-019-0590-x

ORIGINAL ARTICLE

Social BI to understand the debate on vaccines on the Web and social media: unraveling the anti-, free, and pro-vax communities in Italy

Matteo Francia[1] · Enrico Gallinucci[1] · Matteo Golfarelli[1]

Received: 25 February 2019 / Revised: 14 June 2019 / Accepted: 31 July 2019 / Published online: 10 August 2019

Abstract

The debate on vaccines in Italy has greatly intensified in recent years. The promulgation of a law that makes a set of ten vaccines obligatory has pushed this formerly niche topic to a mainstream level. The law itself is an answer to the progressive erosion of the vaccine coverage. The debate has become a political topic with three main positions: supporters of the importance of vaccines, opponents who claim that vaccines are harmful to health, and the new position of those contesting only the mandatoriness of vaccinations. In this paper, we build on a Social Business Intelligence architecture to propose an in-depth analysis of the emerging social debate. Our analysis spans over more than three years, covering all the Web and social media. We adopt several techniques, including community detection and text analytics, to understand the evolution of the debate, the discussed topics, and the structure and peculiarities of the main social communities. The study reveals that the communities are well characterized, especially from a political perspective, and provides useful insights to official health organizations to improve their communication strategies.

Keywords Social Business Intelligence · Vaccines · Community detection · Social networks

1 Introduction

The topic of vaccination has attracted considerable controversy over the years. Since the very beginning of widespread vaccination in the early 1800s, there have been groups of people actively opposing their inoculation (Wolfe and Sharp 2002). Their resistance (which sometimes escalated into riots) leads to the foundation of opposition movements: the "Anti-Vaccination League" was first founded in London in 1853, followed in 1867 by the "Anti-Compulsory Vaccination League," which only focused on the infringement of personal liberty and choice due to vaccines being mandatory (Wolfe and Sharp 2002).

As of today, the belief that vaccinations possibly lead to severe adverse effects is still popular. A recent worldwide survey (Larson et al. 2016) shows that despite the majority of the people being ultimately aware of the benefits of vaccination programs, the percentage of the population raising doubts on their safety or actively opposing to their inoculation is significant and is growing. Researches are alarmed by this situation and worried about public health being at risk, as they underline that the loss of public confidence may lead to the decrease in immunization rates and to the resurgence of diseases (Bonhoeffer and Heininger 2007; Chen 1999). Official health organizations further back this alarm, as they have registered outbreaks of infectious diseases (e.g., measles) in recent years in both Europe and the USA (Zipprich et al. 2015; World Health Organization et al. 2018).

In Italy, the debate on vaccinations experienced has intensified significantly in the last few years. The decreasing vaccine coverage rates and the re-emerging measles outbreaks threatened to compromise the so-called *herd immunity* (Signorelli et al. 2017b, a). The Italian parliament reacted by promulgating a law that introduces the obligation of ten mandatory vaccinations for infants and their exclusion from kindergartens if they are not in compliance with all the

✉ Enrico Gallinucci
enrico.gallinucci@unibo.it

Matteo Francia
m.francia@unibo.it

Matteo Golfarelli
matteo.golfarelli@unibo.it

[1] DISI – University of Bologna, Via dell'Universitá 50, 47522 Cesena, Italy

required vaccinations (Donzelli and Demicheli 2018).[1] This resulted in a widespread controversy, as a free-vax movement has emerged to oppose against the new law and contest its compliance with the Italian Constitution.[2] As a result, the media coverage and the public participation on social media on this topic have been particularly wide (Furini and Menegoni 2018).

With the advent of social networks, the debate on vaccinations found fertile ground to grow on—not only in Italy but on a worldwide basis (Bello-Orgaz et al. 2017). Researches recognize that the anti-vaccination groups have successfully exploited the power of the Web 2.0 and proactively push medical institution into providing more comprehensive and convincing information concerning the safety of vaccines (Betsch et al. 2012). It has been proven that the exposure to proper Web-based resources can positively influence parental vaccine behaviors (Glanz et al. 2017; Biasio et al. 2016), but the spreading of fake news can create a great deal of confusion on the subject (Ciampaglia 2018), as confirmed by a recent work that studied the issue of vaccine hesitancy in Italy (Giambi et al. 2018). In any case, social networks are certainly playing a central role in the debate and in the campaigns of activists that either advocate for or oppose to vaccines.

This scenario motivated us to gain a deeper knowledge of this phenomenon. The study that we present in this paper was first funded in 2015 and 2016 by the health department of an Italian region (Veneto) and continued as a long-term independent research work. Here, we adopt a Social Business Intelligence approach (Francia et al. 2016) to systematically monitor the active discussion on the theme of vaccinations in Italy: for a period of over 3 years, we have collected all social media posts, blog entries, and Web pages that mentioned this subject.

In this paper, we start with a general overview of the discussion, analyzing trends, occurrences, and correlations of specific topics. In particular, we rely on domain experts from the health department for the definition of the topics of interest that need to be monitored. Then, we go more in depth to analyze the communities that have been formed on Twitter around the discussion of vaccinations from different perspectives. We adopt two distinct approaches to retrieve both active communities (i.e., users that actively participate in the discussion by commenting on vaccinations) and passive communities (i.e., users that read comments from the most influential ones), and we characterize them based on the emerging opinion class (i.e., either in favor of vaccine, against vaccines, or against the mandatoriness of vaccines). Finally, we infer knowledge of these communities based from both structural and semantic points of view.

Differently from previous works on this subject:

- we adopt an approach that enables a long-term and continuous monitoring of the discussion, and we cover a wide time span on a variety of media types;
- the distinction of active and passive communities enables a deeper understanding of the discussion and of the involved users;
- to best of our knowledge, we are the first to consider the free-vax classification (i.e., those that oppose only to the mandatoriness of vaccines) in the definition of communities. This allows us to gather interesting information on a movement that has grown with the introduction of the mandatoriness law.

The paper is structured as follows. Section 2 describes the related works on the theme of vaccinations, while Sect. 3 discusses the methodology we adopted in this study; the results of the analyses are then presented in Sect. 4. We conclude with final remarks and a discussion of future directions in Sect. 5.

2 Related works

The analytical potential hidden in social data has been proven so far by several studies on different domains [e.g., visual impairment campaign (Al Zayer and Gunes 2018), political polarization (Conover et al. 2011; Francia et al. 2016; Trottier and Fuchs 2014), social behaviors (Holmberg and Thelwall 2014), brand analysis (Ghiassi et al. 2013), just to name a few], reaching the highest attention by the mainstream media and population with the recent Cambridge Analytica scandal (Grassegger and Krogerus 2017). A whole research area (called Social Business Intelligence (Francia et al. 2014), or SBI) is dedicated to the integration of the social data flow into the enterprise Business Intelligence pipeline to improve the analytical potential of managers and decision makers (Gallinucci et al. 2015). Thus, it is no surprise that many studies have relied on social data to better understand the discussion on vaccinations and to acquire the necessary knowledge to improve the efficacy of official communication channels (Radzikowski et al. 2016).

A summary of the related work is given in Table 1, where the last line represents our work. Since most research is focused on Twitter, we discuss Twitter-based works in Sect. 2.1 and the others in Sect. 2.2.

[1] Preliminary results on the first years after the promulgation of the law show an increase in vaccination coverage, suggesting a positive influence of said regulation (D'Ancona et al. 2018; Signorelli et al. 2018)

[2] In November 2017, the Italian Constitutional Court has ruled in favor of the law on mandatory vaccinations, declaring it to be legitimate (Petrarca et al. 2018; Ministero della Salute 2017).

Table 1 Summary of related works

Work	Topic	Language	Months	Media type	Sentiment analysis	Community detection	Text analytics
D'Andrea et al. (2017)	Vaccines	IT	3	TW	A	–	–
Dunn et al. (2015)	HPV	EN	6	TW	A	–	–
Kadam (2017)	Vaccines	EN	5	TW	A	–	–
Mitra et al. (2016)	Vaccines	EN	6	TW	A	–	✓
Salathé and Khandelwal (2011)	A(H1N1)	EN	6	TW	A	FW	–
Surian et al. (2016)	HPV	EN	24	TW	M	FW	✓
Yuan and Crooks (2018)	MMR	EN	2	TW	A	RT	–
Bello-Orgaz et al. (2017)	Vaccines	EN	7	TW	M	RT	–
Kang et al. (2017)	Vaccines	EN	2	TW	M	–	✓
Radzikowski et al. (2016)	Vaccines	EN	2	TW	M	–	✓
Furini and Menegoni (2018)	Vaccines	IT	24	FB	M	–	✓
Faasse et al. (2016)	Vaccines	EN	–	FB	M	–	✓
Covolo et al. (2017)	Vaccines	IT	16	YT	M	–	–
Larson et al. (2013)	Vaccines	EN	12	W	M	–	–
Yom-Tov and Fernandez-Luque (2014)	MMR	EN	6	SQ	M	–	–
Our	Vaccines	IT	38	W&S	M	FW	✓

Media type: FB, Facebook; SQ, search queries; TW, Twitter; W, Web pages; W&S, Web and social media; YT, YouTube. Sentiment analysis: A, automatic; M, manual. Community detection: FW, followers; RT, retweets

2.1 Works on Twitter

Among Twitter-based studies, the most popular activity is sentiment analysis, which consists in detecting the polarization (either positive, neutral or negative) expressed in a post by its author; this is usually done by relying on natural language processing (NLP) techniques that infer the sentiment by analyzing the morphology, syntax, and semantics of the text. Although several tools exist for automatic detection of sentiment, related works build their own classifier, train it on a manually labeled sample, and use it to label the whole corpus. D'Andrea et al. (2017) focus on Italian discussion and project the obtained sentiment values on a timeline to understand whether specific events (e.g., the cancelation of the "Vaxxed" movie projection at the Senate) caused a spike in either negative or positive tweets. Dunn et al. (2015) also focus on sentiment trends to demonstrate that the likelihood of a user posting a negative tweet after exposure to a majority of negative opinions is higher than those exposed to a majority of positive and neutral tweets. Other works (Kadam 2017; Mitra et al. 2016; Salathé and Khandelwal 2011) rely on labeled tweets as a starting point to infer the polarization of users and to carry out further analyses. Kadam (2017) runs text analytics on the groups of polarized users and builds a classifier to predict a user's orientation based on the content of her tweets. Mitra et al. (2016) further distinguishes between active users that steadily tweet either in favor or against vaccinations and "sleeping" users that tweet against vaccinations only after a significant event. Linguistic inquiry and word count (LIWC) (Tausczik and Pennebaker 2010) and meaning extraction method (Chung and Pennebaker 2008) text analysis techniques are used to understand social and behavioral characteristics of each group, concluding for instance that anti-vaccination groups tend toward categorical thinking and conspiratorial worldviews, or that they show a higher group cohesion. Salathé and Khandelwal (2011) used the polarization of users to infer the polarization of communities. Starting from the dataset of tweets (focused on the influenza A(H1N1) vaccine delivered in 2009), they build the network of the tweeting users and of their following relationships and run the Spin Glass community detection algorithm (Reichardt and Bornholdt 2006) to discover the existing communities. Ultimately, they conclude that almost all of the significant communities are polarized and that users tend to seek information from those sharing the same opinion (a behavior that increases the risk of an outbreak among nonprotected individuals).

Community detection based on the following relationships is also done by Surian et al. (2016), which uses Louvain (Blondel et al. 2008) and InfoMap (Rosvall and Bergstrom 2008) community detection algorithms and focuses on tweets about the human papillomavirus (HPV) from 2013 to 2015. Here, instead of sentiment analysis, they use latent Dirichlet allocation (LDA) (Blei et al. 2003) and Dirichlet mixture model (DMM) (Nigam et al. 2000) text analysis techniques to obtain (and manually label as positive or negative) the topics mentioned in the tweets; then, they infer the polarization of communities by verifying the prevalence of either kind of topics. Other than the typical pro- and anti-communities, they discover an “experiential” community, consisting of young people tweeting about their experience with vaccines which are at greater risk of exposure to safety concerns.

Other works (Yuan and Crooks 2018; Bello-Orgaz et al. 2017) run community detection algorithms on the network based on retweeting relationships (i.e., a user is linked to another if the first retweeted a tweet from the second one) rather than following relationships [although Surian et al. (2016) claim that the latter yields the highest levels of performance]. Similarly to Salathé and Khandelwal (2011), Yuan and Crooks (2018) use the sentiment detected on tweets to infer the polarization of the communities, identified with the Louvain algorithm (Blondel et al. 2008). By analyzing the communities, they discover that analysis shows that people in the anti-vaccination community tend to communicate between themselves, while those in the pro-vaccination one tend to retweet and respond to anti-vaccination tweets—thus suggesting that studies evaluating the *influence* of pro-vaccination users should consider whether their tweets actually penetrate the anti-vaccination community. Bello-Orgaz et al. (2017) run a comparative evaluation of community detection algorithms (concluding that Fast-Greedy (Clauset et al. 2004) is the one yielding the most cohesive and dense communities) and infer the polarization of communities by manually evaluating the most frequent retweets in the community. Their conclusions are that anti-vaccination communities tend to be less in number and less connected than pro-vaccination ones and that famous people tend to fall in the latter groups.

A different approach is taken by Kang et al. (2017), which focused on the links to external Web sites posted in the tweets: they identify the most popular links, manually label them as positive or negative based on their actual content and transcribe the latter into subject–predicate–object triples, so as to create a semantic network of concepts. Their results show that the pro-vaccination group is more cohesive in terms of the topics and discourses appearing in the linked Web sites.

Radzikowski et al. (2016) focuses instead on hashtags, discovering the most tweeted ones and building a network of their co-occurrences. Their findings show that news stories about health issues are the ones driving public participation, but official public health agencies are not strongly featured in the retweet narrative. In particular, bottom-up campaigns and grass-roots activism far outweigh the impact of top-down efforts from authoritative sources such as CDC and WHO, thus suggesting that an indirect approach from mainstream media may be more effective than a direct approach from governmental agencies.

2.2 Works on other media sources

Although Twitter captures most of the researchers’ attention, there exist works based on different sources. Furini and Menegoni (2018) rely on Facebook to analyze the discussion in Italy: the texts of posts from over 200,000 groups (manually labeled as anti- or pro-) are analyzed to identify the presence of words belonging to some predefined categories (affective, social, medial and biological) and subcategories. Their results indicate that pro-vaccination groups show more anxiety and talk more about health in general, whereas anti-vaccination groups show more anger and talk more about vaccine damages to the human body. Another work on Facebook is Faasse et al. (2016), which uses LIWC to analyze the responses to a popular post on vaccines. Its results indicate that despite the absence of scientific evidence, anti-vaccination people show greater analytical thinking; this suggests that their stance is not “universally critical” of vaccines and that it may originate from differing understandings of the benefits and risks of vaccination.

Ultimately, few works fall outside the scope of social networks. Covolo et al. (2017) focus on Italian YouTube videos to discover that anti-vaccination videos are the most shared and liked ones despite being fewer in number. Larson et al. (2013) suggest an architecture for real-time statistical analysis of content from a variety of sources, including articles, blog posts, and news stories, in order to tailor more effective and timely strategies for immunization programs. Although interesting, their work still relies greatly on manual tasks. Yom-Tov and Fernandez-Luque (2014) take a distinguishing approach and focus on search queries on Bing to understand the behavior of pro- and anti-vaccination users. Findings show that, when looking for information on vaccinations, users on both sides are biased in issuing their queries and that although they ultimately look for the same kind of information, their browsing behavior reacts differently after reading the same content. Interestingly, the reading of extreme opinions always proved to have repulsive effects, steering the user’s browsing behavior to more moderate and less opinionated content.

3 Methodology

An overview of the approach we adopted is provided in Fig. 1. The core part is compliant with the SBI methodology presented in Francia et al. (2014) and is central to the analyses described in this paper. In particular, we distinguish between the *topic-driven* analysis, carried out on the contents crawled in the observed period, and the *network-driven* analysis, which focuses on the communities that have formed around the topic of vaccines.

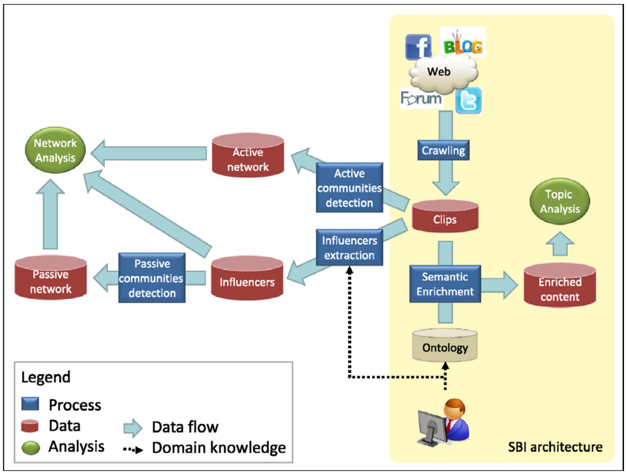

Fig. 1 Approach overview

3.1 The SBI core

The first step in the SBI methodology is the crawling process, required to populate our database with clips collected from the Web. With the term *clip*, we refer to any kind of textual content that comes in the form of a Facebook post, a tweet, a comment on a blog or a news Web site, an article, etc. The retrieval of clips requires the adoption of a certain technique to filter only those related to the specific topic of interest. To this end, *argumentation mining* has emerged as an interesting approach to automatically extract structured arguments from textual documents (e.g., Lippi and Torroni 2016; Habernal and Gurevych 2017). In this research project, we relied on the commercial software Brandwatch (www.brandwatch.com), i.e., a social media monitoring tool for the analysis of user-generated content. It provides a keyword-based crawling service to capture the clips whose content is matched by a certain set of queries. In particular, we issued a single query looking for Italian clips containing any conjugations of the term "vaccine." The query included filters to exclude false positive: most importantly, it discards any mention of animal-related vaccinations and it excludes Web sites known for simply re-posting content from other Web sites.

Clips are downloaded from Brandwatch on a daily basis. As they are stored in the database, a preprocessing step called *semantic enrichment* is taken to run a preliminary textual analysis on the clips' content. The goal is to dissect the textual content into single words so that the ones that have been used the most can be detected. Consistently with the SBI methodology (Francia et al. 2014), we consulted with experts in the local health department of the Italian region of Veneto to define an ontology of topics that they felt relevant to be monitored; we refer to it as the *Topic Ontology*. We started by identifying a set of themes of interest to the experts (e.g., vaccine types, fears, subjects), and we defined the list of topics to be monitored (e.g., "trivalent," "autism," "medic"); each topic comprises a list of different aliases that could be used (e.g., "MMR" and "doctor" are aliases of "trivalent" and "medic," respectively). The Topic Ontology has been maintained throughout the years; Table 2 shows the overall themes identified and the number of defined topics, including some examples. The semantic enrichment process starts by processing the textual content with the open source Apache Lucene library and reducing them to their lemmas (i.e., their canonical form), in order to resolve plurals, masculine and feminine variations; stop words are ignored, while hashtags (i.e., words beginning with the "#" character) are not lemmatized. Then, words are linked to the topics in the Topic Ontology.

Table 2 Themes and topics related to vaccinations, identified with domain experts; the table reports English translations of the Italian terms

Theme	# topics	Example topics
Disinformation	8	Fake news, Hoax, Conspiracy
Fear	18	Aluminum, Autism, Mercury
Influencer	26	Wakefield, Comilva, Burioni
Institution	17	CDC, WHO, ISS, AIFA
Obligatoriness	3	Mandatory, Recommended, Optional
Pharmaceutical house	11	Baxter, Glaxo, Novartis
Scientific term	38	Subcutaneous, Prophylaxis, Drug resistance
Subject	8	Children, Parent, Elderly
Symptom	15	Fever, Rash, Vomit
Television program	5	Le Iene, Openspace, Piazzapulita
Unvaccinable disease	4	AIDS, Ebola, Malaria
Vaccinable disease	27	Measles, Rubella, Tetanus
Vaccine	206	Diftavax, Hexavac, Perturix
Vaccine type	65	DTPA, Hexavalent, Trivalent
Total	451	

An overview of the collected dataset is given in Sect. 4.1.

3.2 Topic-driven analysis

The topic-driven analysis takes place on the semantically enriched content and is aimed at understanding which topics and events are driving the discussions around vaccines. This is mainly done on the topics of the Topic Ontology and on hashtags, which are keywords that often describe synthetically the discussed topic and, possibly, even the opinion of the user. A powerful analysis can be done by examining the *co-occurrence* of different topics among the clips; we say the two topics co-occur in a clip if they are both mentioned in its textual content with no more than 20 words in between them. Co-occurrences can highlight the correlations between different topics: for instance, a public institution can improve the efficiency of its informative campaigns by identifying the kinds of vaccine whose safety is frequently questioned; this can be done by analyzing the co-occurrences between the topics of the Vaccines and Fears theme.

The results of the topic-driven analysis are provided in Sect. 4.2.

3.3 Network-driven analysis

While the topic-driven analysis allows for a general overview of what the conversation around vaccines is about, the network-driven one aims at profiling the people directly or indirectly involved.

3.3.1 Identification and classification of influencers

The core aspect of the network-driven analysis is the identification among the reference population of groups of users that either commonly speak against or in favor of vaccinations, or that are passively subject to certain kinds of opinions. Thus, one of the main aspects consists in labeling users and groups based on their respective opinions. In particular, we discriminate between the following *opinion classes*:

- P: *Pro-vax*, i.e., those speaking in favor of vaccines;
- A: *Anti-vax*, i.e., those speaking against vaccines;
- F: *Free-vax*, i.e., those fighting against the obligation imposed on the children going to kindergartens, but not fighting against vaccines themselves.[3]

In the *influencers extraction* step, we define a *ground-truth* comprising the 50 most influential users for each opinion class. To determine the influence of a user, we rely on the influence measure by Bakshy et al. (2011) (we will refer to it as the Bakshy measure in the remainder of the paper): given a certain user, the influence is computed as a function of the number of her followers and the average number of retweets by her direct followers. Since the Bakshy measure returns discrete values, we introduce a further measure to sort users with the same influence. In particular, we calculate the vaccine follower ratio as the fraction of the user's followers who belong to our reference dataset. In other words, we state that the higher the fraction of followers who tweet about vaccines, the higher is the user's importance in the vaccine domain. The classification of users is done by three domain experts with a majority voting system. Given the list of users ordered by descending influence, each expert manually ascertains their opinion until 50 influencers have been determined for each opinion class. We will use I to refer to the set of influencers and I_A, I_F, I_P to the subsets of influencers in the respective opinion classes.

For the purpose of this analysis, we focus on Twitter alone, as its API service allows for easily obtaining the following relationships between users. (Also, we will show in Sect. 4.1 that Twitter is the most relevant source for vaccine-related discussions.) Since Twitter's APIs return only current data (e.g., the list of a user's followers cannot be filtered on a certain date), we limit the network-driven analysis to the last month of our dataset (i.e., August 2018).

A discussion on the classified influencers is given in Sect. 4.3.1.

3.3.2 Active and passive networks

The goal of the network-driven analysis is to understand the characteristics of two kinds of networks: the *active network*, which comprises people actively discussing about vaccines (because they have tweeted about it at least once in the observed period), and the *passive network*, which comprised people that simply listen to the subject (because they follow at least one influencer). We will refer to users in the active and passive networks as *writers* and *readers*, respectively. In both cases, we aim at grouping people into communities (based on their opinion class) and to evaluate the latter from a structural and a semantic perspective. We will use W and R to refer to communities in the active and passive networks, respectively.

In regard to the active network, we start from the set of people who tweeted about vaccines , collect their relationships through the Twitter APIs (as this information is not provided by our crawling service) and run the well-known Louvain algorithm (Blondel et al. 2008) to detect communities. The Louvain algorithm works similarly to hierarchical clustering, i.e., it groups nodes (and subsequently groups) bottom-up; the goal is to locally maximize the modularity,

[3] Whereas no-vax users are also against the obligation, we distinguish free-vax users for being either in favor of or neutral to vaccinations.

which measures the density of edges inside communities to edges outside communities. Once the communities W are detected, we use the subset of classified users I to verify whether the opinion classes are well partitioned. Ultimately, the claim is that studying the relationships between the users will reveal their opinion.

As to the passive network, we start from the classified users I and study the characteristics of their neighborhood. In particular, we build a community for each opinion class (i.e., $R = \{R_A, R_F, R_P\}$) by considering the set of followers of the respective influencers. (Again, we rely on the Twitter APIs to collect the list of the influencers' followers and the relationships among themselves.) The underlying assumption is that the act of following users with certain opinions augments the probability of the follower to agree with the opinion of the followee. Passive communities differ from active ones in two aspects. First, their cardinality is fixed, i.e., $|R| = 3$, whereas $|W|$ depends on the structure of the network itself and on the algorithm used for community detection. Also, passive communities are overlapped, as users are free to follow influencers from different opinion classes. Thus, we will make reference to the *exclusive communities* R_{A*}, R_{F*}, and R_{P*} (i.e., including users that follow influencers from only one opinion class) and to the *intersection communities* R_{AF}, R_{FP}, R_{AP}, and R_{AFP} (i.e., including users that follow influencers from more than one opinion class).

A discussion on the characteristics of the active and passive networks is given in Sects. 4.3.2 and 4.3.3, respectively.

3.3.3 Structural and semantic analyses of the networks

The analysis of the two networks begins with the validation of W by framing the manually ascertained influencers I. The goal is to verify whether the opinion of the influencers can generalize the opinion of the writers. Then, we compare the communities of the two networks from both structural and semantic points of views. In the first case, we look for similarities and differences between communities in terms of different structural metrics, including their size, cohesion, and overlap degree. In the second case, we are interested in understanding which are the main topics of interest in each community, without restricting the scope to the theme of vaccinations. To this end, we exploit Twitter APIs one last time to collect the most recent tweets (up to 3200, according to Twitter's policy) and we examine the usage that communities make of hashtags (i.e., words preceded by the "#" character), which are frequently used on Twitter to provide the context of the tweet. On the one hand, we calculate the Ruzicka index (Deza and Deza 2006) [i.e., a weighted version of the Jaccard index (Jaccard 1901)] to calculate the pairwise similarity of the communities in terms of the used hashtags; the intuition is that the more two communities use the same hashtags, the more they are similar. On the other hand, we isolate the most distinguishing hashtags for each community and we manually evaluate the existence of any pattern characterizing the communities.

The task of finding distinguishing hashtags resembles the one of finding the most relevant terms within a corpus of documents in information retrieval. Among the most popular techniques to address this task are TF–IDF (term frequency–inverse document frequency) (Salton and McGill 1984) and PMI (pointwise mutual information) (Church and Hanks 1990). The first one multiplies the frequency of a term (TF) by its inverse document frequency (IDF), which quantifies the specificity of a term as an inverse function of the number of documents in which it occurs. The second one quantifies the discrepancy between the probability of the coincidence of two variables (a term and a document) given their joint distribution and their individual distributions, assuming independence. The problem with TF-IDF is that it excludes terms that appear in every document. This works well in a scenario with many documents in order to find *rare* terms that identify some documents; however, in our case, we want to retrieve *popular* hashtags that are mainly used (but not exclusive to) by a certain community. PMI partly addresses this issue by not excluding the latter hashtags, but still risks to confer a higher score to those that are exclusive to a community, independently of their absolute frequency. For instance, a hashtag that appears ten out of ten times in a certain community has a higher PMI score than a hashtag that appears 999 out of 1000 times in the same community.

To isolate the distinguishing hashtags, we introduce a variation of TF-IDF that best addresses our scenario. Given a set of communities C, for each hashtag h used by a community $c \in C$ we compute the term frequency tf(h, c) as in TF-IDF as

$$\mathrm{tf}(h, c) = \frac{f_{h,c}}{f_c}$$

where $f_{h,c}$ is the number of times h has been used by the users in c, and f_c is total number of times any hashtag has been used by the users in c. Then, we compute the term frequency ratio tfr(h, c) as

$$\mathrm{tfr}(h, c) = \frac{\mathrm{tf}(h, c)}{\sum_{c' \in C} \mathrm{tf}(h, c')}$$

In practice, the term frequency ratio measures the relevance of each hashtag in each community based on the calculated term frequencies. Ultimately, we say that a hashtag h is distinguishing of a community c if $\mathrm{tfr}(h, c) > 0.5$ (which is possible for only one community, if any). For the sake of this analysis, we focus on the hashtags that show the highest term frequencies and that satisfy the previous condition.

The results of the structural and semantic comparisons are provided in Sects. 4.3.4 and 4.3.5, respectively. In the

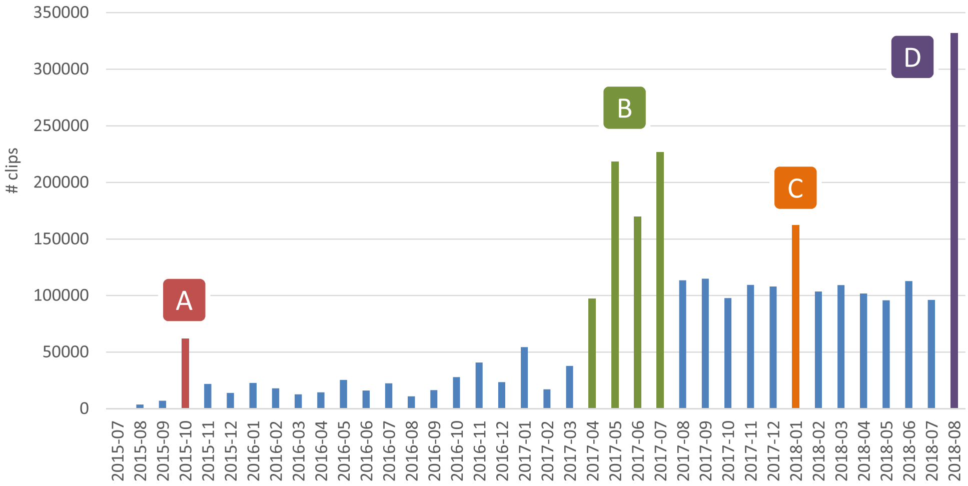

Fig. 2 Monthly trend of clips retrieved on vaccinations; peak periods (letters A to D) are highlighted with different colors and discussed in Sect. 4.1

latter, both PMI and our new approach are used to evaluate the distinguishing hashtags.

4 Results

We first provide a description of our dataset in Sect. 4.1 and then show the results of the topic-driven and network-driven analyses in Sects. 4.2 and 4.3, respectively.

4.1 Dataset description

Our dataset covers a period of over 3 years, spanning from October 2015 to August 2018 and containing a total of 2,207,167 *clips*. Figure 2 shows the monthly trend of clips over the listening period. There are several peak periods that are highlighted in figure, namely:

(A) October 2015 is due to first warnings from the World Health Organization (WHO) concerning the low coverage for infant's vaccinations in Italy.
(B) In April 2017 begin the talks of the mandatory regulation, which is approved in July 2017.
(C) The spike in January 2018 is caused by political talks hinting to a possible weakening of the mandatory regulation.
(D) August 2018 is due to reactions on the promulgation of a law (coming from the newer government) that suspended the mandatoriness for kindergarten enrollments.

Among these events, the second one (B) is the most critical one, as it triggers a significant increase in the talks about vaccinations that becomes structural. Before that time, vaccinations were a niche topic: the discussion was scarce (around 700 posts/day on average) and characterized by occasional and short-lasting spikes (no more than 6000 posts/day) due to some news published by media (e.g., the death of a child, the release of the Vaxxed movie, public statements by important politicians or influencers). After that, the talks on the mandatory regulation pushed the topic to a higher level, as it became a political issue frequently debated by the parties and echoed by the media. Indeed, the discussion has increased by an order of magnitude, with more than 5000 posts/day and spikes of around 50,000 posts/day.

Table 3 shows the distribution of clips across different media channels. Twitter emerges as the most relevant channel despite holding a low market share in Italy (Statista 2018b); this is most probably due to its default policy of making tweets public (unless specified otherwise by the user), whereas the discussion on Facebook [which still retains the highest market share (Statista 2018a)] often takes place within private profiles or protected groups.

4.2 Topic analysis

Figure 3 gives a summary of the monthly presence of the Topic Ontology's themes over the whole dataset; the boxplot also describes the (in)stability of each theme, as the amount of mentions depends on the trending discussions among

Table 3 Distribution of the collected clips on the different media channels

Media channel	# clips
Blog	35,450
Facebook	165,717
Forum	25,743
General	46,688
Image	167
Instagram	3193
News	101,866
Review	55
Twitter	1,231,663
Video	3325

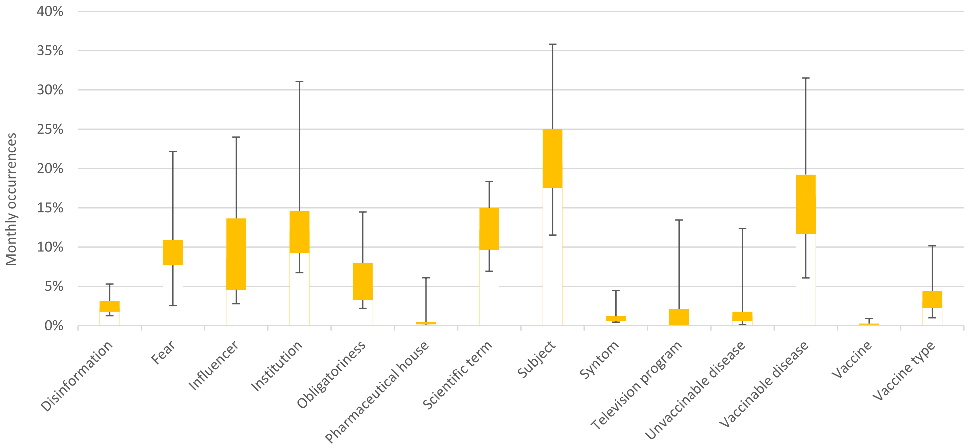

Fig. 3 Boxplot of the monthly occurrences of topics in the Topic Ontology over the whole dataset, grouped by theme

the public opinion. From this analysis, it emerges that subjects and vaccinable diseases are the most present themes. The discussions often revolve around subjects like children (i.e., the main recipients of vaccines), doctors (i.e., those in charge of prescriptions), and parents (i.e., those responsible for the children's health), all of which appear among the ten most mentioned topics (Table 4). Also, vaccinable diseases like measles and meningococcus are very frequent. The first is debated from all sides, as pro-vaccination people claim it should have been eradicated, anti-vaccination ones argue against the actual dangerousness of the disease, and official health organizations often report its death count and outbreaks. The second is debated for being a dangerous disease which has been excluded from the list of mandatory ones. Other important topics include the obligatoriness of vaccines (which jumped to the top 10 only in May 2017), schools (due to the proposal of exclude nonvaccinated children from primary school and kindergartens), politicians and political parties, in particular Beatrice Lorenzin being the Ministry of Health promulgating the law on obligatoriness, and Matteo Salvini (i.e., the leader of "La Lega") and M5S (i.e., Five Stars Movement) as the most popular representatives criticizing the law.

A more in-depth analysis is provided in Fig. 4, which shows the co-occurrences between the topics of themes Vaccine Type (left) and Fear (right); it is a chord diagram generated with D3 (http://d3js.org/), where the size of each chord depends on the number of co-occurrences. Although few correlations cross the boundary between the two themes, interesting conclusions can be derived from this chart. For starters, it is clear that the fear of autism is associated with the trivalent and hexavalent vaccines. This can be a hint for official health organizations to improve the efficiency of their informative campaigns by focusing their efforts on promoting the safety of those two kinds of vaccines. Also, the topic of epidemics is often associated with the anti-poliomyelitis and anti-influenza vaccines; however, this correlation does not indicate an actual fear of the vaccines: a closer look to the involved clips reveals that the anti-poliomyelitis is cited for being capable of stopping one of the latest great outbreaks, while the anti-influenza is advertised to contain the epidemic that spreads every year. Still on the anti-influenza, its correlation with pregnancy is due to the former one being particularly recommended to weak or exposed subjects, such as pregnant women. Ultimately, it is worth mentioning that the absence of any correlation from mercury and aluminum to any of the vaccine types suggests that the fear of the former is more generic; for this reason, official health organizations may need to design a communicative strategy that tackles these topics in a different way than with autism.

4.3 Network analysis

The results of the network analysis are provided in Sects. 4.3.1–4.3.5. To simplify the reading, we summarize in Table 5 the symbols used to reference networks and communities.

Table 4 Ten most mentioned topics, their total number of occurrences in the dataset and the % of clips in which they occur at least once

Topic	Theme	Occurrences	% of clips
Child	Subject	343,658	13
Mandatory	Obligatoriness	200,812	9
Doctor	Subject	161,714	6
School	Institution	140,346	5
M5S	Influencer	133,411	5
Measles	Vaccinable disease	101,720	4
Matteo Salvini	Influencer	89,261	3
Beatrice Lorenzin	Influencer	85,407	4
Meningococcus	Vaccinable disease	66,660	3
Parent	Subject	65,841	3

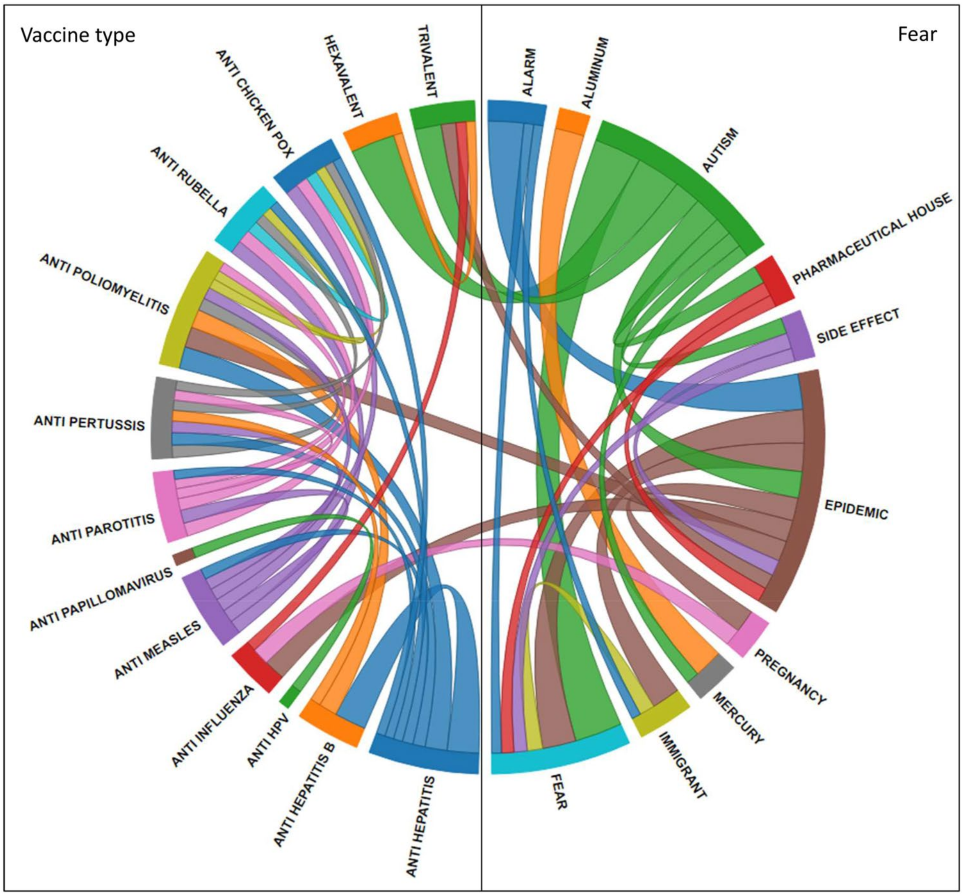

Fig. 4 Co-occurrences between the topics of themes Vaccine Type (left) and Fear (right)

Table 5 Summary of the symbols used to reference networks and communities

Symbol	Description
I	Set of manually classified influencers
I_A, I_F, I_P	Subset of manually classified influencers of a certain opinion class
W	Active network, i.e., the writers
W_{AF}, W_{P_i}, W_{U_i}	Active network communities, labeled with the representative opinion classes (U = unclassified)
R	Passive network, i.e., the readers
R_A, R_F, R_P	Overlapping communities in the passive network, labeled with respect to the opinion class of the followed influencers
R_{A*}, R_{F*}, R_{P*}	Exclusive communities in the passive network, labeled with respect to the opinion class of the followed influencers
$R_{AF}, R_{FP}, R_{AP}, R_{AFP}$	Intersection communities in the passive network, labeled with respect to the opinion class of the followed influencers

Table 6 Influence level of the top 50 influencers manually identified for each opinion class

Infl. level	Bakshy range	I_A	I_F	I_P
High	[90, 100)	6	23	50
Medium	[40, 90)	1	4	–
Low	[20, 40)	27	23	–
Very low	[1, 20)	16	–	–

4.3.1 The classified influencers

A summary of the manually classified users I is shown in Table 6, where users are grouped into four levels based on their influence value of the Bakshy measure: high, medium, low or very low. This reveals a substantial difference between the three classes, which resembles the observations made in Bello-Orgaz et al. (2017). On the one side, the

Table 7 Statistics on the 20 communities detected in the active network, including the number of users, the percentage with respect to the total, and the number of enclosed anti-, free-, and pro-vax influencers

$w \in W$	$\lvert w \rvert$	$\frac{\lvert w \rvert}{\lvert W \rvert}$ (%)	$\lvert w \cap I_A \rvert$	$\lvert w \cap I_F \rvert$	$\lvert w \cap I_P \rvert$
W_{AF}	8506	13	50	48	–
W_{P_1}	11,850	18	–	1	3
W_{P_2}	29,727	45	–	1	11
W_{P_3}	8624	13	–	–	36
W_{U_1}	7668	12	–	–	–
W_{U_2}	173	0	–	–	–
Others	270	0	–	–	–

most important anti-vax influencers users are not particularly influencing, as the vast majority of them detain the lowest levels of influence; their set is mainly made of ordinary people, and there is not a single celebrity or famous person among them. On the opposite side, all pro-vax influencers fall within the highest level influence; here, not only ordinary people can be found, but also well-known journalists, politicians, and medics. In particular, the "front man" is Roberto Burioni (Wikipedia 2019), an immunologist that has become famous for his battle in favor of vaccines. Interestingly, pro-vax influencers also include several satirical blogs. The free-vax influencers occupy the middle ground, both in terms of influence level (almost perfectly balanced) and number of famous people, mainly bloggers and politicians. Another interesting evaluation can be done from the political perspective: free-vax influencers mainly fall within the sphere of the Five Star Movement, pro-vax ones mainly belong to the left-wing side, and anti-vax ones hardly show any political affiliation.[4]

4.3.2 The active network

Over the period observed in this analysis (i.e., August 2018), the dataset presents a total of 66,645 users that had tweeted on the topic of vaccinations. The application of Louvain's community detection algorithm on these users produces a total of 20 communities with a modularity value[5] of 0.4154. Details of the communities are reported in Table 7, while Fig. 5 provides a graphical representation (result obtained using the igraph Python library, https://igraph.org/python/); in the figure, colored squares represent the classified users.

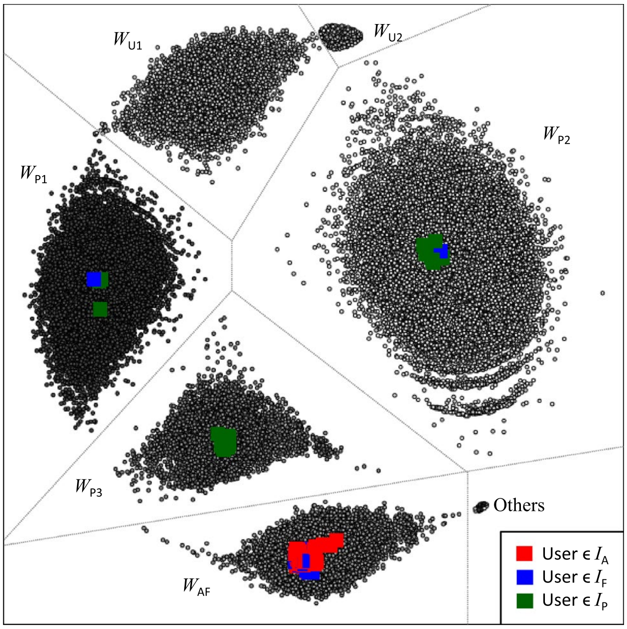

Fig. 5 Graphical representation of the communities detected in the active network; dots represent users, and colored squares represent users ∈ I

Interestingly, the major communities that arise provide a clear partitioning of the influencers based on their classification; thus, we have named the communities based on the opinion class of the contained influencers (U stands for unclassified). All anti-vax influencers and almost all free-vax influencers are concentrated in W_{AF}; conversely, pro-vax influencers characterize three different communities, namely W_{P_2} (the biggest one, containing the influencers with the highest media exposure, including Roberto Burioni), W_{P_3} (a strong community including the majority of the influencers), and W_{P_1} (a weaker community, including only few influencers). These statistics alone suggest that:

1. Pro-vax users are higher in number and spread in different communities.
2. There is a net separation between the anti-vax and the pro-vax world.
3. The free-vax world shares much more with the anti-vax world than with the pro-vax one.

In the remainder of the paper, we will use W_P to refer to $\{W_{P_1} \cup W_{P_2} \cup W_{P_3}\}$.

4.3.3 The passive network

Starting from I, we have collected all their followers R ($|R| = 472{,}047$) and all the following relationships that occur between them. These users are assigned to R_A, R_F, and R_P depending on whether they follow some of the

[4] Political orientation has been identified through a manual approach: for public figures with verified accounts, we used the declared profile; for other, users we browsed their tweets and inferred the political orientation from their content.

[5] Modularity ranges from − 1 to 1, where a positive number indicates a dense network.

Table 8 Statistics on the communities detected in the passive network, including the number of users and the percentage with respect to the total

$r \in R$	Formula	$\lvert r\rvert$	$\frac{\lvert r\rvert}{\lvert R\rvert}$ (%)
R_A	R_A	90,063	19
R_F	R_F	294,858	62
R_P	R_P	218,665	46
R_{A*}	$R_A \setminus \{R_F \cup R_P\}$	35,291	7
R_{F*}	$R_F \setminus \{R_A \cup R_P\}$	215,105	46
R_{P*}	$R_P \setminus \{R_A \cup R_F\}$	138,912	29
R_{AF}	$R_A \cap R_F$	36,915	8
R_{FP}	$R_F \cap R_P$	27,967	6
R_{AP}	$R_P \cap R_A$	2986	1
R_{AFP}	$R_A \cap R_F \cap R_P$	14,871	3
Total	$R_A \cup R_F \cup R_P$	472,047	100

classified users in the respective class. Details of each community are reported in Table 8, while Fig. 6 provides a graphical representation; in the figure, each circle corresponds to a community in Table 8 and its size is proportional to the number of users. These data show that:

1. There is little overlapping between the pro-vax and anti-vax communities, i.e., very few people are following both pro- and anti-vax influencers. This suggests that the campaign of pro-vax influencers in favor of vaccinations is probably going to have little impact on the anti-vax community.
2. The anti-vax community is actually small if compared to the others, which is due to the absence of truly influential anti-vax users.
3. An unexpected result is the greater size of the free-vax community with respect to the pro-vax one. This may be explained by the higher presence of Five Star Movement politicians among the top free-vax influencer compared to the presence of left-party politicians among the top pro-vax influencers.

4.3.4 Structural comparison of the two networks

A comparison of the active network W and the passive network R is shown in Table 9. For each community c, we show the number of users ($|c|$), the number of following relationships between the users of such community (intra(c)), and two measures of cohesion (Wasserman and Faust 1994), i.e., the density (dens(c)), and the average centrality degree (centr(c)). Density is calculated as the number of relationships against all possible relationships (i.e., $\mathrm{dens}(c) = \frac{\mathrm{intra}(c)}{|c|*|c-1|}$), while the average centrality degree is the average number of relationships for each user $u \in c$ such that u follows another $u' \in c$. The table also shows the degree of overlap between the communities: for each passive community c, it shows the percentage of writers from either W_{AF} or W_P that also belongs to c. Conversely, the last column shows the percentage of users of c that are also writers. We summarize our findings as follows.

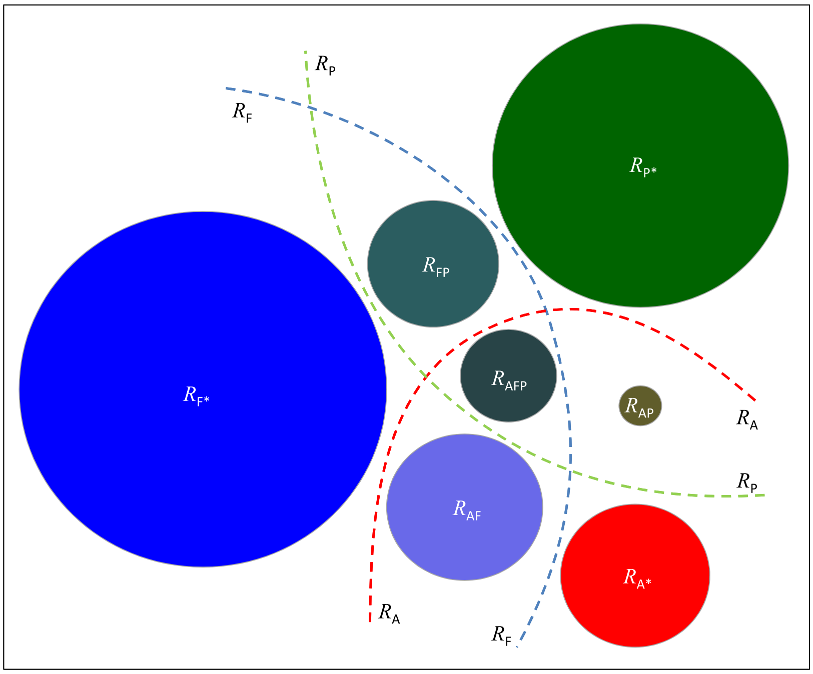

Fig. 6 Graphical representation of the communities detected in the passive network

Table 9 Metrics to compare the detected communities, including the number of users (|c|), the number of following relationships such as users (intra(c)), the density (dens(c)), the average centrality degree (centr(c)), the percentage of writers from either W_{AF} or W_{P} that also belongs to c, and the percentage of users of c that are also writers

$c \in \{W, R\}$	|c|	intra(c)	dens(c)	centr(c)	$\frac{\lvert c \cap W_{\mathrm{AF}}\rvert}{\lvert W_{\mathrm{AF}}\rvert}$ (%)	$\frac{\lvert c \cap W_{\mathrm{P}}\rvert}{\lvert W_{\mathrm{P}}\rvert}$ (%)	$\frac{\lvert c \cap W\rvert}{\lvert c\rvert}$ (%)
W_{P_1}	11,850	732,248	5.2	62	0	24	100
W_{P_2}	29,727	1,758,127	2.0	59	0	59	100
W_{P_3}	8624	1,008,819	13.6	117	0	17	100
W_{AF}	8506	1,086,976	15.0	128	100	0	100
R_{A}	90,063	17,768,812	2.2	197	65	5	9
R_{F}	294,858	47,512,996	0.5	161	82	20	6
R_{P}	184,736	30,444,425	0.9	165	26	61	18
$R_{\mathrm{A}*}$	35,291	447,025	0.4	13	2	0	1
$R_{\mathrm{F}*}$	215,105	5,804,724	0.1	27	17	5	2
$R_{\mathrm{P}*}$	138,912	5,557,235	0.3	40	1	46	17
R_{AF}	36,915	2,543,931	1.9	69	41	1	11
R_{FP}	27,967	2,230,534	2.9	80	3	11	21
R_{AP}	2986	37,259	4.2	12	0	1	15
R_{AFP}	14,871	4,567,752	20.7	307	22	3	23

All density values (dens(c)) are multiplied by 10^{-3}

1. We note that 89% of writers are also present as readers; this means that the vast majority of users tweeting on vaccinations are also followers of at least one of the top 150 influencers.
2. In terms of density, we observe that the cohesion of the active communities is higher than in passive ones; this seems intuitive, as writers are actively talking about the subject of this study, while readers are mostly spectators of the debate (as shown by the overlapping percentages). Interestingly, the most cohesive active communities are W_4 and W_{10}, i.e., those that contain the highest number of influencers (as seen in Table 7).
3. When looking at the exclusive and overlapping passive communities, we notice that the cohesion measures progressively increase when moving from the former to the latter. Indeed, R_{AFP} is the one with the highest density and average centrality degree, even with respect to the active communities. This can be explained by the fact that the more users are open to interacting with influencers of different sides, the more interconnected they become.
4. The overlapping percentages seem to confirm the classification of the communities, as 46% of users in W_{P} fall within $R_{\mathrm{P}*}$ and 41% of users in the W_{AF} fall within R_{AF}. This also strengthens the suggestion that there is a high affiliation between anti-vax and free-vax users.

4.3.5 Semantic comparison of the two networks

Besides comparing communities from a structural perspective, we conclude by comparing them from a semantic point of view. The selection of the most distinguishing hashtags (based on both PMI and our approach, presented in Sect. 3.3.3) depends on which communities are taken into account. Thus, given C the set of considered communities, we evaluate the four scenarios described in Table 10. Scenarios S$_1$ and S$_2$ differ in that the three pro-vax active communities are considered separately and together, respectively; scenario S$_3$ considers the three big overlapping passive communities, while S$_4$ considers the exclusive and the intersection passive communities.

The tables presenting the results are available at Mendeley Data (http://dx.doi.org/10.17632/2hfw99xz44.2). As anticipated in Sect. 3.3.3, PMI is more biased toward the exclusivity of hashtags in a community (to the detriment of the hashtags' frequency), whereas our approach brings out popular hashtags that are more used by a certain community. Nonetheless, both techniques consistently return results that support the same conclusion in most scenarios. In the discussion of our findings, we first comment the results obtained with our technique and then compare them with those obtained with PMI.

Overall, the presented scenarios reveal that one of the main topics of discussion (besides vaccines) is politics, especially in the active community. Before commenting on the results, we clarify that, at the time of the analysis, the ruling government is the one formed on June 1, 2018 by

Table 10 Different scenarios considered for the selection of the most distinguishing hashtags; the results are publicly available at Mendeley Data (http://dx.doi.org/10.17632/2hfw99xz44.1)

Scenario	Set of communities C
S$_1$	$\{W_{\mathrm{AF}}, W_{\mathrm{P}_1}, W_{\mathrm{P}_2}, W_{\mathrm{P}_3}\}$
S$_2$	$\{W_{\mathrm{AF}}, \{W_{\mathrm{P}_1} \cup W_{\mathrm{P}_2} \cup W_{\mathrm{P}_3}\}\}$
S$_3$	$\{R_{\mathrm{A}}, R_{\mathrm{F}}, R_{\mathrm{P}}\}$
S$_4$	$\{R_{\mathrm{A}*}, R_{\mathrm{F}*}, R_{\mathrm{P}*}, R_{\mathrm{AF}}, R_{\mathrm{FP}}, R_{\mathrm{AP}}, R_{\mathrm{AFP}}\}$

Table 11 Pairwise Ruzicka similarities between communities

	I_A	I_F	I_P	W_{AF}	W_{P_1}	W_{P_2}	W_{P_3}	R_A	R_F	R_P
I_A	–	0.51	0.11	0.47	0.13	0.14	0.13	0.04	0.03	0.03
I_F	0.51	–	0.12	0.52	0.14	0.14	0.13	0.02	0.01	0.01
I_P	0.11	0.12	–	0.15	0.45	0.46	0.58	0.01	0.01	0.01
W_{AF}	0.47	0.52	0.15	–	0.19	0.20	0.18	0.06	0.06	0.05
W_{P_1}	0.13	0.14	0.45	0.19	–	0.50	0.55	0.04	0.03	0.03
W_{P_2}	0.14	0.14	0.46	0.20	0.50	–	0.62	0.05	0.05	0.05
W_{P_3}	0.13	0.13	0.58	0.18	0.55	0.62	–	0.05	0.05	0.04
R_A	0.04	0.02	0.01	0.06	0.04	0.05	0.05	–	0.61	0.48
R_F	0.03	0.01	0.01	0.06	0.03	0.05	0.05	0.61	–	0.62
R_P	0.03	0.01	0.01	0.05	0.03	0.05	0.04	0.48	0.62	–

the Five Star Movement (5SM, a big-tent party) and Lega (a right-wing party); these are the parties that have been most critical on the obligatoriness of vaccines, which had been promulgated in 2017 by the Democratic Party (DP, a left-wing party). We summarize our findings as follows.

1. S_1 reveals a clear characterization of the active communities. Writers in W_{AF} not only show support for the free-vax movement, but they also attack the former Minister of Health and former Prime Minister that promulgated the law on vaccine obligatoriness (i.e., Beatrice Lorenzin and Matteo Renzi, respectively) and their political party (i.e., DP). Conversely, writers in W_{P_2} (i.e., the largest community) are the ones showing the strongest support for the pro-vax movement, while those in W_{P_3} (i.e., the one with the majority of influencers) are mainly critical of the current government; writers in W_{P_1} (i.e., the weaker pro-vax community) cover a middle ground (showing pro-vax support and critics to the government) and often cite satirical Web sites. Hashtags retrieved by PMI confirm these considerations.
2. S_2 confirms the clear separation between W_{AF} and the pro-vax communities. On the one side, the distinguishing hashtags for W_{AF} are mostly the same as in S_1. On the other side, the distinguishing hashtags that emerge for W_P can be traced back to the same topics, i.e., pro-vax support and critics to the government and to free-vax politicians. As in S_1, the same is evident in the PMI results.
3. S_3 shows other themes besides politics that characterize the passive communities. Readers in R_A show support toward the Lega party and interest in political talk shows and finance topics; news events appear as the main topic among readers R_F, while those in R_P mostly discuss reality shows. These considerations are less apparent in the PMI results, where only the interest in news events appears in R_F.
4. S_4 shows that the support toward the Lega party and the interest in talk shows can be narrowed down to R_{AF}, while the interest in finance topics can be narrowed down to R_{A*}. Beyond that, no other main topic seems to characterize the remaining communities. Similarly to S_3, no main topic of interest is highlighted by PMI.
5. From a linguistic perspective, a surprising result is the higher tendency of the distinct groups to argue against those with opposing views rather than directly supporting the respective view. For instance, in both PS_1 and S_2 pro-vax politicians and parties are attacked by the anti- and free-vax communities, while free-vax politicians and parties are attacked by the pro-vax communities. This is probably a symptomatic effect of the social network dialectic, in which it is easier to criticize "enemies" rather than supporting "friends."
6. The terms "pro-vax" and "no-vax" (the latter referring to the anti-vax movement) appear as distinguishing hashtags in W_{AF} and W_P, respectively. This shows that such terms are mainly used as pejorative labels by the respective sides rather than for self-identification.

We conclude the analysis by discussing the pairwise Ruzicka similarity between communities in terms of the most used hashtags; the intuition is that the more two communities use the same hashtags, the more they are similar. The results are presented in Table 11. The comparison is also made with the top influencers in each opinion class, i.e., I_A, I_F, and I_P. We summarize our findings as follows.

1. Among the influencers (which are strongly focused on the theme of vaccination), we observe a high similarity between the anti-vax and the free-vax (which further confirms the strong connection between the two groups), whereas both of them appear quite distant from the pro-vax.

2. This trend is further confirmed in the comparison of active communities: the anti-vax community (i.e., W_4) shares similarities only with I_A and I_F, while all pro-vax communities share similarities with each other and with I_P.
3. As to the passive communities, we primarily focus on the exclusive ones (i.e., R_A, R_F, and R_P). Users in these communities are less focused on the theme of vaccination and more active on the mainstream topics; as a result, the widespread usage of mainstream hashtags makes interest areas appearing quite similar to each other and, at the same time, quite far from every influencer group and every active community. Nonetheless, it is still observable how users in R_F actually cover a middle ground of topics between those in R_A and R_P.

5 Conclusion

In this paper, we analyzed the social debate on vaccines in Italy. The topic is relevant since in the last few years it passed from being a niche audience to a mainstream argument due to the promulgation of a law that makes a set of ten vaccines obligatory that turned vaccine in a political issue. Differently from previous papers, we integrated several techniques and we considered all the available Web and social sources. This allowed us to carry out a comprehensive analysis that tackled the topic from several different points of view. The main outcomes of the analysis are: (a) the fear of autism is mostly associated with the trivalent and hexavalent vaccines; (b) the anti-vax community is quite smaller than the pro-vax one but is also more cohesive; (c) there is no public figure among the anti-vax influencers, contrarily to the pro- and free-vax ones; (d) the free-vax community shares many similarities with the anti-vax one; and (e) there is a clear characterization of the active communities, especially from a political perspective. Finally, as computer science researchers, we deliberately avoid an evaluation from the sociological and medical perspectives, and we advocate for professionals in these fields to exploit the facts and evidences that have emerged in this study.

We will continue monitoring the vaccine topic with particular interest to the evolution of the network structure and to the topic discussed. In particular, it will be interesting to analyze how emerging positions and opinions (like the one experienced in 2017) impact on the Twitter follower network. This requires the adoption of a dynamic approach that allows the tracking of the evolution of communities and the association of such changes to significant events (Wang et al. 2018). Remarkably, since Twitter APIs do not allow the retrieval of historical relationships (nor do they associate a time reference with such relationships), this is possible only by continuously monitoring the relationships in the community.

References

Al Zayer M, Gunes MH (2018) Exploring visual impairment awareness campaigns on twitter. Soc Netw Anal Min 8(1):40

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining. ACM, pp 65–74

Bello-Orgaz G, Hernandez-Castro J, Camacho D (2017) Detecting discussion communities on vaccination in twitter. Future Gener Comput Syst 66:125–136

Betsch C, Brewer NT, Brocard P, Davies P, Gaissmaier W, Haase N, Leask J, Renkewitz F, Renner B, Reyna VF et al (2012) Opportunities and challenges of web 2.0 for vaccination decisions. Vaccine 30(25):3727–3733

Biasio LR, Corsello G, Costantino C, Fara GM, Giammanco G, Signorelli C, Vecchio D, Vitale F (2016) Communication about vaccination: a shared responsibility. Hum Vaccines Immunother 12(11):2984–2987

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan):993–1022

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008

Bonhoeffer J, Heininger U (2007) Adverse events following immunization: perception and evidence. Curr Opin Infect Dis 20(3):237–246

Chen RT (1999) Vaccine risks: real, perceived and unknown. Vaccine 17:S41–S46

Chung CK, Pennebaker JW (2008) Revealing dimensions of thinking in open-ended self-descriptions: an automated meaning extraction method for natural language. J Res Personal 42(1):96–132

Church KW, Hanks P (1990) Word association norms, mutual information, and lexicography. Comput Linguist 16(1):22–29

Ciampaglia GL (2018) Fighting fake news: a role for computational social science in the fight against digital misinformation. J Comput Soc Sci 1(1):147–153

Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066111

Conover M, Ratkiewicz J, Francisco MR, Gonçalves B, Menczer F, Flammini A (2011) Political polarization on twitter. Icwsm 133:89–96

Covolo L, Ceretti E, Passeri C, Boletti M, Gelatti U (2017) What arguments on vaccinations run through youtube videos in italy? A content analysis. Hum Vacc Immunother 13(7):1693–1699

D'Ancona F, D'Amario C, Maraglino F, Rezza G, Ricciardi W, Iannazzo S (2018) Introduction of new and reinforcement of existing compulsory vaccinations in Italy: first evaluation of the impact on vaccination coverage in 2017. Eurosurveillance 23(22):1800238

D'Andrea E, Ducange P, Marcelloni F (2017) Monitoring negative opinion about vaccines from tweets analysis. In: 2017 third international conference on research in computational intelligence and communication networks (ICRCICN). IEEE, pp 186–191

Deza E, Deza M (2006) Dictionary of distances. North-Holland, Amsterdam

Donzelli A, Demicheli V (2018) Varicella vaccination: scientific reasons for a different strategic approach. Epidemiol Prev 42(1):65–70

Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E (2015) Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. J Med Internet Res 17(6):e144

Faasse K, Chatman CJ, Martin LR (2016) A comparison of language use in pro-and anti-vaccination comments in response to a high profile facebook post. Vaccine 34(47):5808–5814

Francia M, Golfarelli M, Rizzi S (2014) A methodology for social bi. In: Proceedings of the 18th international database engineering & applications symposium. ACM, pp 207–216

Francia M, Gallinucci E, Golfarelli M, Rizzi S (2016) Social business intelligence in action. In: International conference on advanced information systems engineering. Springer, pp 33–48

Furini M, Menegoni G (2018) Public health and social media: language analysis of vaccine conversations. In: 2018 international workshop on social sensing (SocialSens). IEEE, pp 50–55

Gallinucci E, Golfarelli M, Rizzi S (2015) Advanced topic modeling for social business intelligence. Inf Syst 53:87–106

Ghiassi M, Skinner J, Zimbra D (2013) Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Expert Syst Appl 40(16):6266–6282

Giambi C, Fabiani M, D'Ancona F, Ferrara L, Fiacchini D, Gallo T, Martinelli D, Pascucci MG, Prato R, Filia A et al (2018) Parental vaccine hesitancy in Italy-results from a national survey. Vaccine 36(6):779–787

Glanz JM, Wagner NM, Narwaney KJ, Kraus CR, Shoup JA, Xu S, O'Leary ST, Omer SB, Gleason KS, Daley MF (2017) Web-based social media intervention to increase vaccine acceptance: a randomized controlled trial. Pediatrics 140(6):e20171117

Grassegger H, Krogerus M (2017) The data that turned the world upside down. Vice Motherboard. https://www.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win. Accessed 4 Jan 2019

Habernal I, Gurevych I (2017) Argumentation mining in user-generated web discourse. Comput Linguist 43(1):125–179. https://doi.org/10.1162/COLI_a_00276

Holmberg K, Thelwall M (2014) Disciplinary differences in twitter scholarly communication. Scientometrics 101(2):1027–1042

Jaccard P (1901) Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaud Sci Nat 37:547–579

Kadam M (2017) Understanding vaccination attitudes and detecting sentiment stimulus in online social media. PhD thesis, Illinois Institute of Technology

Kang GJ, Ewing-Nelson SR, Mackey L, Schlitt JT, Marathe A, Abbas KM, Swarup S (2017) Semantic network analysis of vaccine sentiment in online social media. Vaccine 35(29):3621–3638

Larson HJ, Smith DM, Paterson P, Cumming M, Eckersberger E, Freifeld CC, Ghinai I, Jarrett C, Paushter L, Brownstein JS et al (2013) Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines. Lancet Infect Dis 13(7):606–613

Larson HJ, de Figueiredo A, Xiahong Z, Schulz WS, Verger P, Johnston IG, Cook AR, Jones NS (2016) The state of vaccine confidence 2016: global insights through a 67-country survey. EBioMedicine 12:295–301

Lippi M, Torroni P (2016) Argumentation mining: state of the art and emerging trends. ACM Trans Internet Technol 16(2):10:1–10:25. https://doi.org/10.1145/2850417

Ministero della Salute (2017) Decreto vaccini, la sentenza della corte costituzionale considera legittimo l'obbligo dei vaccini nel contesto attuale. http://www.salute.gov.it/portale/news/p3_2_1_1_1.jsp?lingua=italiano&menu=notizie&p=dalministero&id=3184. Accessed 04-June-2019

Mitra T, Counts S, Pennebaker JW (2016) Understanding anti-vaccination attitudes in social media. In: ICWSM, pp 269–278

Nigam K, McCallum AK, Thrun S, Mitchell T (2000) Text classification from labeled and unlabeled documents using EM. Mach Learn 39(2–3):103–134

Petrarca L, Midulla F, Openshaw PJ (2018) Vaccination policies in europe: common goals, diverse approaches and public doubts. Eur J Immunol 48(1):10–12

Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL (2016) The measles vaccination narrative in twitter: a quantitative analysis. JMIR Public Health Surveill 2(1):e1

Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74(1):016110

Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105(4):1118–1123

Salathé M, Khandelwal S (2011) Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 7(10):e1002199

Salton G, McGill M (1984) Introduction to modern information retrieval. McGraw-Hill Book Company, New York

Signorelli C, Guerra R, Siliquini R, Ricciardi W (2017a) Italy's response to vaccine hesitancy: an innovative and cost effective national immunization plan based on scientific evidence. Vaccine 35(33):4057–9

Signorelli C, Odone A, Cella P, Iannazzo S, D'Ancona F, Guerra R (2017b) Infant immunization coverage in italy (2000–2015). Annali dell'Istituto superiore di sanita 53(3):231–237

Signorelli C, Iannazzo S, Odone A (2018) The imperative of vaccination put into practice. Lancet Infect Dis 18(1):26–27

Statista (2018a) Social network market share held by Facebook in Italy from January 2017 to August 2018. https://www.statista.com/statistics/622874/facebook-s-social-network-market-share-monthly-in-italy/

Statista (2018b) Social network market share held by Twitter in Italy from January 2017 to August 2018. https://www.statista.com/statistics/622878/twitter-s-social-network-market-share-monthly-in-italy/

Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG (2016) Characterizing twitter discussions about HPV vaccines using topic modeling and community detection. J Med Internet Res 18(8):e232

Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 29(1):24–54

Trottier D, Fuchs C (2014) Social media, politics and the state: protests, revolutions, riots, crime and policing in the age of facebook, twitter and youtube, vol 16. Routledge, Abingdon

Wang Z, Li Z, Yuan G, Sun Y, Rui X, Xiang X (2018) Tracking the evolution of overlapping communities in dynamic social networks. Knowl Based Syst 157:81–97. https://doi.org/10.1016/j.knosys.2018.05.026

Wasserman S, Faust K (1994) Social network analysis: methods and applications, vol 8. Cambridge University Press, Cambridge

Wikipedia (2019) Roberto burioni. http://en.wikipedia.org/wiki/Roberto_Burioni. Accessed 04 Jan 2019

Wolfe RM, Sharp LK (2002) Anti-vaccinationists past and present. Br Med J: BMJ 325(7361):430

World Health Organization et al (2018) Europe observes a 4 fold increase in measles cases in 2017 compared to a previous year. World Health Organization 6

Yom-Tov E, Fernandez-Luque L (2014) Information is in the eye of the beholder: seeking information on the MMR vaccine through an internet search engine. In: AMIA annual symposium proceedings, vol 2014. American Medical Informatics Association, p 1238

Yuan X, Crooks AT (2018) Examining online vaccination discussion and communities in twitter. In: Proceedings of the 9th international conference on social media and society. ACM, pp 197–206

Zipprich J, Winter K, Hacker J, Xia D, Watt J, Harriman K (2015) Measles outbreak-california, December 2014–February 2015. MMWR Morb Mortal Wkly Rep 64(6):153–154

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.